Haplotype tagging reveals parallel formation of hybrid races in two butterfly species

Joana I. Meier[a,b,1], Patricio A. Salazar[a,c,1], Marek Kučka[d,1], Robert William Davies[e], Andreea Dréau[d], Ismael Aldás[f], Olivia Box Power[a], Nicola J. Nadeau[c], Jon R. Bridle[g], Campbell Rolian[h], Nicholas H. Barton[i], W. Owen McMillan[j], Chris D. Jiggins[a,j,2,3], and Yingguang Frank Chan[d,2,3]

[a]Department of Zoology, University of Cambridge, Cambridge CB2 3EJ, United Kingdom; [b]St. John's College, University of Cambridge, Cambridge CB2 1TP, United Kingdom; [c]Department of Animal and Plant Sciences, University of Sheffield, Sheffield S10 2TN, United Kingdom; [d]Friedrich Miescher Laboratory of the Max Planck Society, 72076 Tübingen, Germany; [e]Department of Statistics, University of Oxford, Oxford OX1 3LB, United Kingdom; [f]Baños, Ecuador; [g]Department of Genetics, Evolution and Environment, University College London, London WC1E 6BT, United Kingdom; [h]Department of Comparative Biology and Experimental Medicine, Faculty of Veterinary Medicine, University of Calgary, Calgary, AB T2N 4N1, Canada; [i]Institute of Science and Technology Austria, 3400 Klosterneuburg, Austria; and [j]Smithsonian Tropical Research Institute, Panamá, Apartado Postal 0843-00153, República de Panamá

Edited by Molly Przeworski, Columbia University, New York, NY, and approved May 5, 2021 (received for review July 20, 2020)

Genetic variation segregates as linked sets of variants or haplotypes. Haplotypes and linkage are central to genetics and underpin virtually all genetic and selection analysis. Yet, genomic data often omit haplotype information due to constraints in sequencing technologies. Here, we present "haplotagging," a simple, low-cost linked-read sequencing technique that allows sequencing of hundreds of individuals while retaining linkage information. We apply haplotagging to construct megabase-size haplotypes for over 600 individual butterflies (*Heliconius erato* and *H. melpomene*), which form overlapping hybrid zones across an elevational gradient in Ecuador. Haplotagging identifies loci controlling distinctive high- and lowland wing color patterns. Divergent haplotypes are found at the same major loci in both species, while chromosome rearrangements show no parallelism. Remarkably, in both species, the geographic clines for the major wing-pattern loci are displaced by 18 km, leading to the rise of a novel hybrid morph in the center of the hybrid zone. We propose that shared warning signaling (Müllerian mimicry) may couple the cline shifts seen in both species and facilitate the parallel coemergence of a novel hybrid morph in both comimetic species. Our results show the power of efficient haplotyping methods when combined with large-scale sequencing data from natural populations.

butterfly | genomes | hybrid zone | population genetics | haplotypes

Understanding how changes in DNA sequence affect traits and shape the evolution of populations and species has been a defining goal in genetics and evolution (1–3). DNA is naturally organized in the genome as long molecules consisting of linked chromosome segments. Linkage is a core concept in genetics: in genetic mapping, geneticists map causal variants not by tracking the actual mutation but through many otherwise neutral and unremarkable linked variants. Likewise, the detection of selection relies on observing hitchhiking of linked variants rather than seeing the mutation itself. This recognition makes it all the more paradoxical that haplotype information is routinely omitted from most genomic studies as a technical compromise. Lacking haplotype information not only complicates analysis and ancestry reconstruction but also precludes detection of allele-specific expression (4) and chromosome rearrangements and reduces power to detect selective sweeps, even entirely missing them when multiple haplotypes sweep together (5). Instead of sequencing genomes as haplotypes, short-read sequencing produces 150-bp reads. Until long-read platforms become sufficiently accurate and affordable, this lack of haplotype context will continue to impact mapping and genomic studies, particularly those in nonmodel organisms.

One way to simplify haplotype reconstruction and inference from sequencing data is to avoid discarding haplotype information in the first place. A promising emerging technique is linked-read (LR) sequencing (6–9), which preserves long-range information via molecular barcoding of long DNA molecules before sequencing. Individual short reads can then be linked via a shared barcode to reconstruct the original haplotype. However, existing options all suffer from high cost, poor scalability, and/or require custom sequencing primers or settings that have thus far prevented them from being applied as the default sequencing platform (*SI Appendix*, Tables S1 and S2). If LR sequencing could become scalable and affordable, it would significantly advance genetics by enabling the "haplotyping" of entire populations (i.e., the sequencing and systematic discovery of genomic variants as haplotypes in hundreds or even thousands of samples in model and nonmodel organisms alike).

Here, we describe a solution called "haplotagging," a simple and rapid protocol for LR sequencing. Importantly, haplotagging maintains full compatibility with standard Illumina sequencing

Significance

A defining goal in genetics is linking variation in DNA sequence to trait evolution between populations and, ultimately, species. Genome sequencing efficiently captures such variation but typically in millions of tiny fragments that omit haplotype or linkage information. We present "haplotagging," a simple, rapid linked-read sequencing technique that allows high-throughput sequencing without sacrificing haplotype information. We validated this affordable approach for whole-genome haplotyping in large populations. We used haplotagging to investigate the rise of a novel hybrid morph in parallel hybrid zones of two comimetic *Heliconius* butterfly species in Ecuador. Our results reveal that strikingly parallel divergences in their genomes produced coordinated shifts in haplotype frequencies across the hybrid zone, giving rise to comimetic hybrid morphs in each species.

Author contributions: J.I.M., P.A.S., M.K., C.R., W.O.M., C.D.J., and Y.F.C. designed research; J.I.M., P.A.S., M.K., O.B.P., N.J.N., J.R.B., C.R., W.O.M., C.D.J., and Y.F.C. performed research; J.I.M., P.A.S., M.K., R.W.D., A.D., I.A., O.B.P., N.J.N., J.R.B., C.R., W.O.M., C.D.J., and Y.F.C. contributed new reagents/analytic tools; J.I.M., P.A.S., M.K., R.W.D., N.H.B., C.D.J., and Y.F.C. analyzed data; and J.I.M., N.H.B., C.D.J., and Y.F.C. wrote the paper with contributions from all authors.

Competing interest statement: M.K., A.D., and Y.F.C. declare competing financial interests in the form of patent and employment by the Max Planck Society. The European Research Council provides funding for the research but no other competing interests.

This article is a PNAS Direct Submission.

[1]J.I.M., P.A.S., and M.K. contributed equally to this work.

[2]C.D.J. and Y.F.C. contributed equally to this work.

[3]To whom correspondence may be addressed. Email: c.jiggins@zoo.cam.ac.uk or frank.chan@tue.mpg.de.

This article contains supporting information online at https://www.pnas.org/lookup/suppl/doi:10.1073/pnas.2015005118/-/DCSupplemental.

Published June 21, 2021.

and can easily scale to large populations with no extra costs. We demonstrate this in three steps. First, we show that direct haplotyping using haplotagging is robust in single human and mouse samples with known haplotypes ("phases"). Next, we show the feasibility of population haplotyping in 245 mice, even with very low-coverage LR sequencing. Finally, we apply haplotagging to investigate the emergence of a hybrid morph in a hybrid zone system in Ecuador featuring 670 individuals of two species of *Heliconius* butterflies.

Results

Direct Haplotype Tagging. Haplotagging is a bead-based protocol for the production of LR DNA-sequencing libraries. Haplotagging works via molecular barcoding of long, kilobase-spanning DNA molecules to generate short fragments for sequencing. In solution, DNA molecules tend to wrap around a single bead, a property that can be exploited for constructing LR libraries (8, 9). Each haplotagging bead is tethered with Tn5 transposase carrying one of 85 million molecular barcodes directly integrated into an otherwise standard Nextera Tn5 transposon adaptor (Fig. 1*A* and *SI Appendix*, Fig. S1 and Table S3). In a single transposition reaction, microbead-tethered Tn5 transposase transfers the barcoded sequencing adaptors into the long DNA molecule. A tube of beads carrying millions of unique molecular barcodes can be used to tag a pool of DNA molecules, each carrying its bead-specific barcode ("beadTag," consisting of 4 × 6 nt segments). Following sequencing, unique long-range haplotypes can be reconstructed from each DNA molecule (Fig. 1*A*).

Haplotagging features three main design improvements over other LR alternatives (*SI Appendix*, Table S1). Firstly, it avoids specialized instrumentation (cf. microfluidics chips and controller for 10x Genomics' discontinued Chromium platform). Haplotagging is, in essence, a one-step, 10-min transposition reaction followed by PCR. It requires only a magnet and standard molecular biology equipment available in most laboratories. A haplotagging library, in our hands, costs less than 1% of a 10x Genomics Chromium library and, despite featuring long-range haplotype information, costs about 1/20th as much as a Nextera DNAFlex short-read library (*SI Appendix*, Table S2). Secondly, we designed the segmental beadTag barcode and the protocol with scalability and high-order multiplexing in mind. A single person can thus prepare and sequence hundreds of uniquely barcoded libraries within weeks. Last but not least, one of the major design challenges we have solved in haplotagging is encoding 85 million barcodes and maintaining full compatibility with standard Illumina TruSeq sequencing that is available at most sequencing facilities, even when pooled with other library types (Fig. 1*A*).

To test the recovery of molecular haplotypes, we performed haplotagging on high-molecular-weight DNA from an F1 hybrid mouse between two inbred laboratory strains with known sequence differences, CAST/EiJ (CAST) and C57BL/6N (BL6 is the genome reference strain). We could assign 94.3% of 201 million read pairs to a beadTag and inferred molecules based on beadTag sharing (Fig. 1*B*). Across the genome, we found that 99.97% of phase-informative molecules accurately capture the parental haplotypes with exclusively BL6 (reference, maternal) or CAST (alternate, paternal) alleles at multiple single nucleotide polymorphisms (SNPs) (Fig. 1*B*), with even representation of both alleles (50.6% and 49.4%, respectively; about 1.0 million molecules each on autosomes). Many of these molecules span many kilobases (kbp), up to as much as 415 kbp (N_{50} = 42.1 kbp; *SI Appendix*, Table S4). These results provide strong evidence that haplotagging can accurately capture and reconstruct haplotypes.

Using these data, we performed de novo phasing without prior knowledge and could phase nearly all heterozygous SNPs (99.74%) using HAPCUT2 (10) into large, megabase-spanning phased haplotype blocks (N_{50} = 15.5 Mbp, *SI Appendix*, Figs. S2*A* and S3*A*; maximum: 61.46 Mbp, *SI Appendix*, Table S4; see *SI Appendix*, *SI Text* for phasing performances in additional human and mouse samples, including comparison against other LR platforms).

Whole-Population Haplotyping. We next tackled haplotype phasing using LR data in large populations. Unlike phasing in single individuals, population phasing can be challenging because neither the number of haplotypes nor their frequencies are known in advance. Other studies have yet to apply population phasing using LR data, presumably because such large datasets have not been feasible before.

Our strategy involves leveraging naturally occurring haplotype blocks in populations and trading-off linkage against sequencing (read) coverage: first, we reconstruct the set of haplotypes present in the study population by pooling reads from many individuals for maximum coverage (because most segregating haplotypes are common). Next, we infer or impute the two haplotypes carried by each individual across the genome by evaluating the evidence at the haplotype block level (i.e., using linkage) in addition to raw per-base read coverage. In other words, for phasing purposes, each molecule represents a sampled haplotype and may contribute phasing information along the entire length of the molecule even if very little of each molecule is covered by actual reads (Fig. 1 *B* and *C*). Due to the large molecule size in haplotagging data, each molecule (i.e., linked set of reads) is both more likely to be informative and to often show 10-fold or greater expansion in molecular coverage (Fig. 1*C*) (e.g., in the F1 hybrid mouse); instead of the nominal 12.6× read coverage, there was 165.6× molecular coverage (i.e., each parental haplotype was sampled more than 80 times) (*SI Appendix*, Figs. S2*B* and S3*B* and Table S4). This strategy dovetails neatly with STITCH, an algorithm for (short) read-based statistical phasing (11), which we have adapted to incorporate LR information. The implication of this principle of whole-population phasing is potentially profound: with haplotagging, we can sequence populations at a fraction of current read coverage (and costs), yet still obtain accurate haplotypes for the entire population.

To test this concept, we performed haplotagging on 245 "Longshanks" mice from a 20-generation selective-breeding experiment for long tibia length (12, 13). We sequenced these mice to an average depth of 0.24× and phased the data using STITCH. We tested the accuracy of genotype imputation by comparing against higher-coverage conventional short-read data for 32 of these mice (2.9× coverage) (13). Our results show that genotype imputation using data from haplotagging is remarkably robust (>96% accurate) and remains so even when read coverage is reduced to 0.07× (versus 0.15× without using LR information) (*SI Appendix*, Fig. S3*B* and *SI Text*). In practical study design, these improvements may already be tangible [e.g., halving genotyping costs and improving genotyping accuracy (14), thus improving mapping power]. The most notable improvement, however, is in phasing: compared to short-read sequencing without linked-read data, there was a 100-fold expansion in the length of phased blocks, with an average of 24.1 kbp using LR data (versus 283 bp otherwise). We achieved these robust results despite having sequenced only about 1/10th as deeply [cf. typically 20 to 40× under classical phasing (15) or 2 to 5× when imputing using an external reference haplotype panel (11, 14, 16)]. Importantly, we are no longer dependent on a reference haplotype panel, which does not exist for most study populations.

The exact saving in raw sequencing effort depends on haplotype diversity and structure: the greater the range of linkage disequilibrium (LD) relative to the molecule size, the more efficiently low-coverage sequencing can sample the underlying allele via haplotype tagging. When LD is high (e.g., in selective breeding or admixed populations), haplotype blocks are relatively large, and exhaustive sequencing in each individual may yield only diminishing returns. Conversely, in large, panmictic populations with little LD (<100 bp versus tens of kilobases in an average molecule), the main advantage with LR data would be improved phasing (see *SI Appendix*,

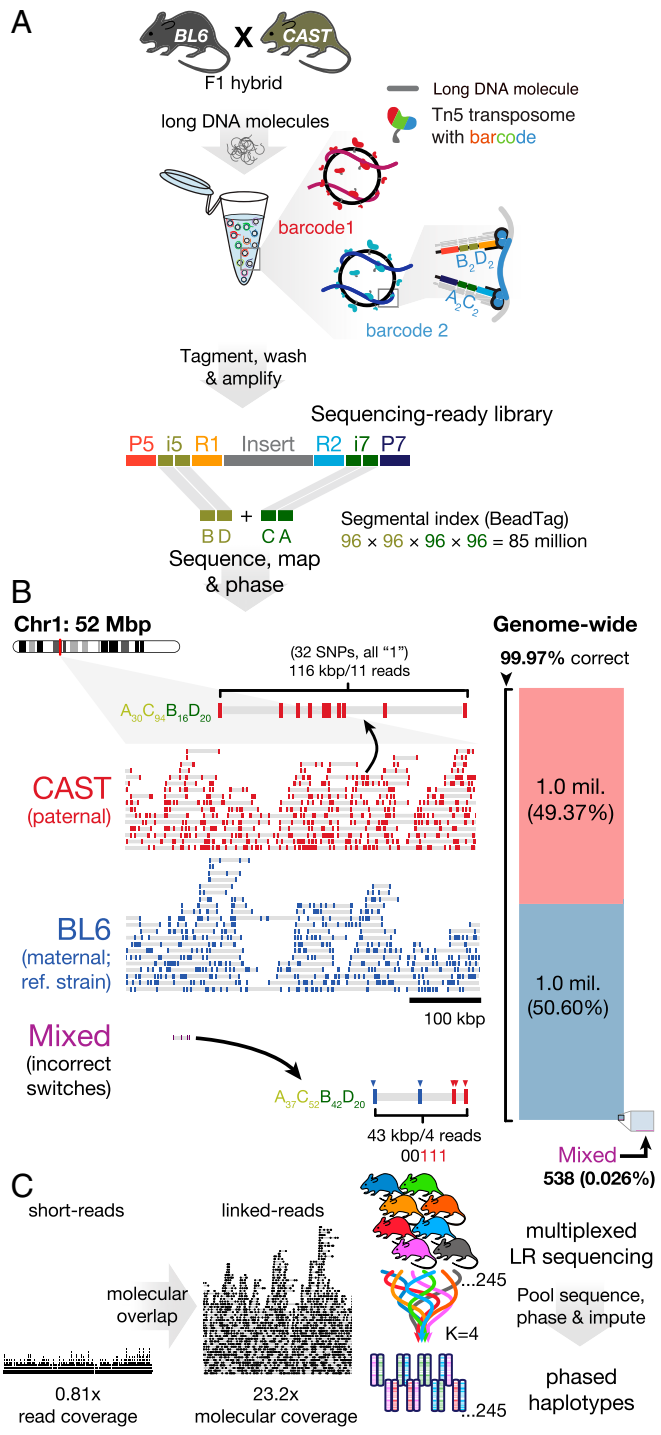

Fig. 1. Haplotagging enables population-scale LR sequencing. (*A*) Principles of haplotagging. Microbeads coated with barcoded transposon adaptors enable simultaneous molecular barcoding and Tn5-mediated fragmentation of long DNA molecules into sequencing-ready libraries after PCR amplification, all in a single tube. This technique takes advantage of the tendency of DNA to interact only with a single bead in solution (*Inset*). A key feature of haplotagging is that each bead is uniformly coated with a single segmental barcode combination ("beadTag") made up of four segments of 96 barcodes each (designated "B," "D," olive and "C," "A," green at the standard i5/7 index positions of the Illumina Nextera design). Across beads, the four segments represent up to 96^4 or 85 million beadTags. Thus, DNA molecules wrapped around a single bead can be reconstructed from individual short reads that share the same beadTag. (*B*) Haplotagging in an F1 hybrid mouse between the reference strain C57BL/6 (BL6) and CAST/EiJ (CAST), with detailed view at Chr1: 52 to 52.5 Mbp. Each molecule is represented by a gray bar connecting short reads (colored bars for CAST, red; or BL6, blue) sharing a single beadTag (e.g., $A_{30}C_{94}B_{16}D_{20}$ tags a 116-kbp molecule carrying a CAST allele). All but one molecule in this window match perfectly to CAST or BL6 alleles. Genome wide, 99.97% of all reconstructed molecules correspond to CAST or BL6 haplotypes (2 million correct versus 538 incorrect molecules). (*C*) Vast expansion in molecular versus read coverage for whole-population haplotyping. LR molecules typically span tens of kilobases, compared to ~500 bp short reads. The increased overlap among molecules often lead to >10-fold increase in molecular coverage (an average of 0.81 reads overlapping a given position versus an average of 23.2 molecules here; *SI Appendix*, Table S4). In a large population, LR data allow both accurate haplotype reconstruction using pooled read depths and accurate imputation by leveraging linkage information, even with input read coverage reduced to 0.07× (*SI Appendix*, Fig. S3*B*). Bead and Tn5 image modified with permission from Zinkia Entertainment, S.A./Pocoyo.

SI Text for detailed discussion). At such a diversity level, it may ultimately require full (LR) sequencing to recover the actual segregating haplotypes in an individual. However, in populations in which LD exceeds molecule size (e.g., recent admixture), researchers can now perform (ultra) low-coverage Illumina sequencing by multiplexing haplotagging libraries of hundreds of samples and leveraging the combined coverage for accurate haplotype reconstruction in each individual.

Parallel *Heliconius* Hybrid Zones. We next applied haplotagging to address key evolutionary questions in two *Heliconius* butterfly species in Ecuador. Local collectors have previously noticed the abundance of a morph of putative hybrid origin in Eastern Ecuador. Here, we investigate the patterns of phenotypic and genetic variation across the hybrid zone where this novel morph is found and test the hypothesis that the morph has arisen and spread in parallel in two species that mimic each other.

Heliconius butterflies have diversified into many species and subspecies (or "races") across South and Central America and represent a classic example of adaptive radiation (17). They are toxic and advertise their unpalatability with bright warning coloration. Predators (mainly birds) learn to avoid the warning signal (18, 19), and selection favors the locally common pattern (20). *Heliconius* species often converge on the same color patterns to reinforce the advertising effect, a phenomenon known as Müllerian mimicry (Fig. 2*A*; ref. 18). The striking variety notwithstanding, the genetics of these color patterns is remarkably simple, involving only a few loci of large effect (17, 21–26).

Here, we focus on two distantly related *Heliconius* species, *Heliconius erato* and *Heliconius melpomene* (diverged 12 million years ago) (27), which feature many distinct color patterns and mimic each other (and other species) whenever they overlap (17). In the Pastaza valley of eastern Ecuador, a highland race of each species meets and forms a hybrid zone with a distinct lowland race (Fig. 2*A* and *SI Appendix*, Fig. S4; refs. 28 and 29). The hybrid zones range from around 1,300 m to 400 m from the Andean mountains into the Amazon basin (highland race: *H. erato notabilis* and *H. melpomene plesseni*; lowland race: *H. erato lativitta* and *H. melpomene malleti*; Fig. 2). To survey the hybrid zone, we collected 975 *H. erato* and 394 *H. melpomene* butterflies (928 and 343 at 35 transect sites; Fig. 2 and *SI Appendix*, Fig. S4 and Tables S6 and S7) and scored their color traits as informed by controlled laboratory crosses (*SI Appendix*, Fig. S5) (28). Fig. 2*B* shows that hybrid butterflies (both F1 and beyond) are observed in all but five highland sites and one lowland site, with the core transition zone between 1,000 and 900 m elevation (kilometers 36 to 45 along the transect).

Divergence, Selection, and Trait Mapping. Using haplotagging, we sequenced 484 *H. erato* and 187 *H. melpomene* butterflies from the transect in 96-plex batches to a median read coverage of 1.29× for

EVOLUTION

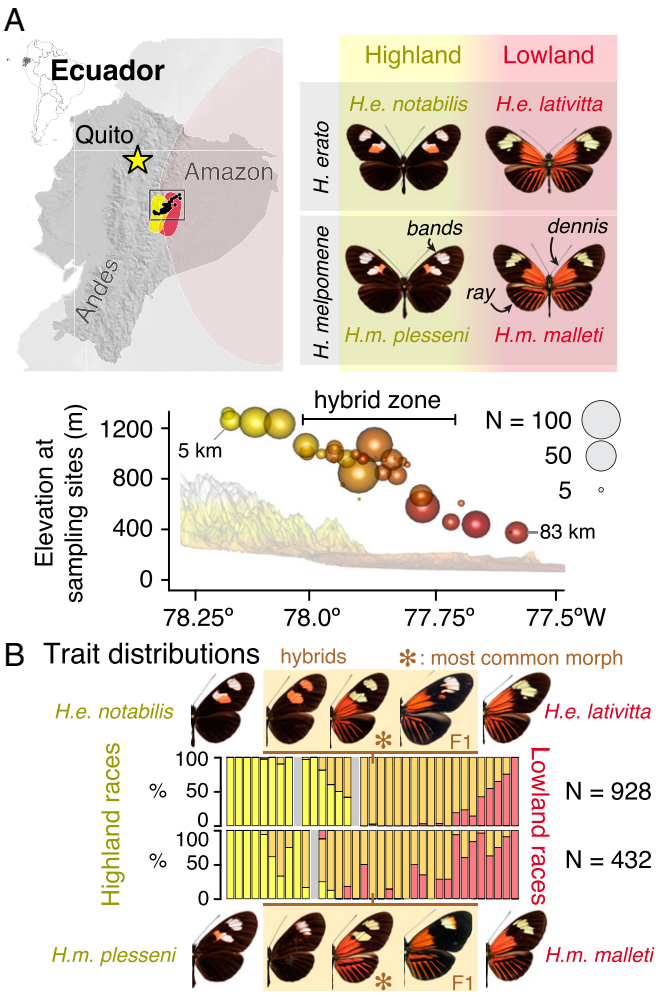

Fig. 2. Parallel hybrid zones in a pair of Müllerian comimicking *Heliconius* butterflies. (*A*) In eastern Ecuador, butterflies of the species *H. erato* and *H. melpomene* occur in the transition zone between the Andes (up to 1,307 m elevation, "Highland") and the Amazon basin (376 m, "Lowland") as distinctive races with major wing color pattern differences (labeled as "bands," "*dennis*," and "*ray*"). *Heliconius* butterflies are unpalatable and share warning wing patterns (Müllerian comimicry) (18). We sampled a total of 1,360 butterflies of both species along an 83-km transect consisting of 35 sampling sites across the double hybrid zones (kilometers 19 to 59; symbols scaled to sample size and colors indicate elevation) and 12 additional off-transect sites (*SI Appendix*, Table S5). (*B*) Proportions of butterflies displaying the highland double-band phenotype (*H. erato notabilis* and *H. melpomene plesseni*: yellow) and lowland *dennis–ray* patterns (*H. erato lativitta* and *H. melpomene malleti*: red) as well as hybrid patterns (F1 and beyond: orange; ✻, most common morph; *SI Appendix*, Fig. S4*B*; gray: sites with no specimen in one species) at sampling sites along the transect.

H. erato and 2.72× for *H. melpomene* (samples both individually and molecularly barcoded; see *Materials and Methods* and *SI Appendix*, Fig. S2*C* and Tables S3 and S5–S8). Following phasing and imputation, we retained a set of 49.2 M SNPs in *H. erato* and 26.3 M SNPs in *H. melpomene*, most of which were polymorphic throughout the zone (*H. erato*: 69.4%; *H. melpomene*: 81.1%), consistent with high gene flow. By contrast, only 232 SNPs were completely fixed for opposite alleles between *H. erato notabilis* and *H. erato lativitta*, and none were fixed between *H. melpomene plesseni* and *H. melpomene malleti*. Sequence diversity was high (131 and 97 SNPs/kbp in *H. erato* and *H. melpomene*, respectively), which, together with molecules spanning tens of kilobases (13.9 kbp and 12.0 kbp on average for *H. erato* and *H. melpomene*, respectively) helped to produce long phased block sizes averaging 3.6 and 3.3 Mbp in *H. erato* and *H. melpomene*, respectively (maximum: 20.7 Mbp, effectively spanning a whole chromosome; *SI Appendix*, Fig. S6). This dataset represents a qualitative jump in quality and resolution over the state-of-the-art in a natural population study.

Across the genome, there was little background genomic differentiation between highland and lowland races in both *H. erato* and *H. melpomene* (mean genetic distance F_{ST} in *H.* erato: 0.0261; in *H. melpomene*: 0.0189; Fig. 3). This is consistent with free introgression of neutral and globally adaptive variants in hybrid zones (30, 31).

Against this backdrop, peaks in genomic differentiation stand out in stark contrast in each species. Using a hidden Markov model (HMM), we identified 24 and 52 regions of high differentiation in *H. melpomene* and *H. erato*, respectively (Fig. 3 and *SI Appendix*, *SI Materials and Methods*). The strongest divergence peaks are found at four major loci, namely *WntA* (Chromosome 10) (23), *vvl* (Chr13), *cortex* (Chr15) (32, 33), and *optix* (Chr18) (17, 22, 34), and they are highly correlated between the two species, highlighting the fine-scale parallelism at these loci (Fig. 3 *B* and *C* and *SI Appendix*, Fig. S7). The improved resolution from haplotagging reveals loci and greater parallelism than previously described (29), with F_{ST} peaks at *cortex* in *H. erato* and *vvl* in *H. melpomene* being the most surprising because they have not previously been thought to play a role in phenotypic divergence in these races (see *SI Appendix*, Fig. S8 *A* and *B* for a direct comparison to ref. 29).

To identify loci controlling wing color patterns, we performed genome-wide association studies (GWAS) and showed that genotypes at *WntA* and *optix* are nearly perfectly concordant for the two most contrasting differences in both species (number of forewing bands: *WntA* on Chr10 and Dennis and Ray red patterns; *optix* on Chr18; likelihood ratio test, $P \ll 1 \times 10^{-150}$ at both loci in *H. erato* and $P < 1 \times 10^{-36}$ in *H. melpomene*; Fig. 3 *B* and *C* and *SI Appendix*, Fig. S5; see *SI Appendix*, *SI Materials and Methods* for details). More subtly, we found *cortex* to control a yellow spot at the forewing base (*cortex: H. erato:* $P \ll 1 \times 10^{-150}$; *H. melpomene:* $P < 1 \times 10^{-10}$; Fig. 3 *B* and *C*). Despite parallel F_{ST} signals, genetic control of wing pattern is not strictly parallel between the species: *vvl* controls forewing band shape only in *H. erato* ($P < 1 \times 10^{-300}$), whereas the same trait is explained by *WntA* in *H. melpomene*. Conversely, the locus containing *cortex*, besides controlling the yellow spot in both species, also acts as a modifier for the distribution of red scales but only in *H. melpomene* ($P < 1 \times 10^{-18}$; Fig. 3*C* and *SI Appendix*, Fig. S5 *F–H*). Together with the wider divergence peak at *cortex* in *H. melpomene* butterflies, it raises the possibility that in *H. melpomene*, *cortex* may be in tight linkage with additional gene(s) and/or regulatory element(s) that control red-scale distribution. Close examination of butterflies carrying recombinant haplotypes shows that the locus in fact consists of at least two tightly linked but distinct genes. The *cortex* gene itself controls the yellow spot (35), with the highland *plesseni*-like allele dominant; in contrast, the lowland, *malleti*-like allele is dominant in controlling the distribution of red scales and maps to *domeless*/*washout* >150 kbp downstream of *cortex* (*SI Appendix*, Fig. S9).

All four loci show reduced nucleotide diversity and elevated LD, characteristic signatures of selective sweeps (captured by the π, ω, and integrated haplotype score (iHS) statistics; Fig. 3 and *SI Appendix*, Fig. S10 and Tables S9 and S10 and Datasets S4 and S5; refs. 36 and 37). Here, the LR data allows us to track the breakdown of haplotypes across the hybrid zone (*SI Appendix*, Fig. S11) and greatly increase the power of the haplotype-based ω-statistic (especially compared to haplotypes reconstructed from short-read only data, *SI Appendix*, Fig. S8*C*). The resolution in these tests is high enough to separate the *dennis* and *ray cis*–regulatory regions from the target gene *optix* in *H. erato* (*SI Appendix*, Fig. S8*C*). The data can reveal unsuspected molecular details as evidenced by the detection of rare *H. melpomene* recombinants at the tightly linked

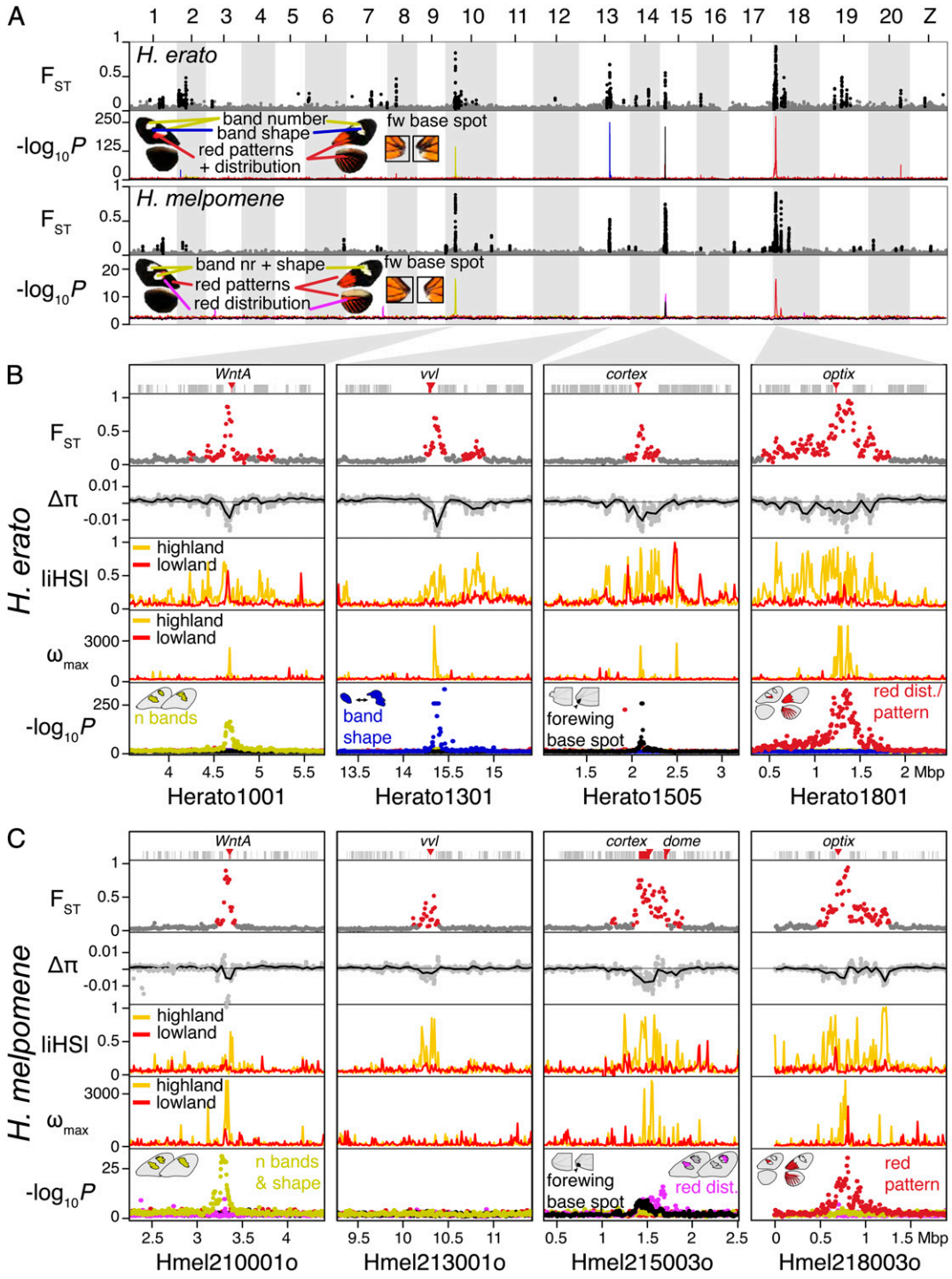

Fig. 3. Highly parallel patterns of differentiation at genomic regions underlying wing color patterns. (*A*) Major peaks of differentiation are shared across *H. erato* and *H. melpomene* (as indicated by F_{ST}; *H. melpomene* data are plotted at its homologous *H. erato* coordinates). F_{ST} values of 10-kbp windows assigned to the high differentiation state by the HMM analysis are shown in black, others in gray. The three most strongly differentiated regions in each pair of subspecies all show strong association with color pattern differences [$-\log_{10}(P)$: $-\log_{10}(P$ value) of the likelihood ratio test, most strongly associated SNP per 10-kbp window shown]. (*B* and *C*) Detailed view at the four loci with strongest differentiation in *H. erato* (*B*) and *H. melpomene* (*C*). At all four major loci, the races also differ in nucleotide diversity (π; $\Delta\pi{=}\pi_{highland} - \pi_{lowland}$), whereby the highland races (*H. erato notabilis* and *H. melpomene plesseni*) consistently show greater reduction in diversity than the lowland races (*H. erato lativitta* and *H. melpomene malleti*), indicative of strongest selection in the highland races in both species. Compared to the $\Delta\pi$ values of all genomic 50-kbp windows, the four major loci are among the most negative 1% in both species (*SI Appendix*, Fig. S10). Stronger selection among highland races than lowland races is also supported by haplotype-based selection statistics such as absolute normalized integrated haplotype score (|iHS|) and the ω-statistic. Three of the four major loci in each species are associated with major color patterns, and all fall into the vicinity of the genes *WntA* (forewing band number), *vvl* (forewing band shape in *H. erato,* ref. 25), *cortex* (yellow spot at the forewing [fw] base in both species and distribution of red scales in *H. melpomene* likely controlled by *domeless*/*washout* [*dome*]), and *optix* (presence of red either as forewing patch and hindwing bar and rays [Dennis–Ray] or in forewing band) in *H. erato* (*B*) and in *H. melpomene* (*C*) (for details see *SI Appendix*, Figs. S9 and S12 and Datasets S4 and S5).

band, *dennis*, and *ray* elements at *optix*. Here, recombinants show that the presence of red scales in the forewing bands is controlled solely by *dennis* but not *band*, as in other *H. melpomene* races (*SI Appendix*, Fig. S12*A*; refs. 38 and 39). In addition, one particularly informative recombinant helps to refine the *ray* element from 14.1 kbp to 8.0 kbp (*SI Appendix*, Fig. S12*B*). Together, our data underscore the precision and power of population haplotyping.

Chromosome Inversions and Other Structural Rearrangements. In local adaptation and speciation, chromosome rearrangements, and inversions in particular, are thought to play a major role in holding together adaptive variants (40–43). However, they are hard to detect using short-read techniques. By contrast, longer LR molecules that span rearranged junctions can systematically reveal insertions, deletions, inversions, and additional chromosome rearrangements. We therefore analyzed beadTag sharing across adjacent 10-kbp windows to detect differences between the physical molecules and the reference assembly (Fig. 4; see *Materials and Methods* for details). We detected 685 and 415 indels and 14 and 19 major inversion/translocation events in *H. erato* and *H. melpomene*, respectively (Datasets S2 and S3).

Although structural rearrangements occasionally overlap divergent peaks or signatures of selection in single high- or lowland populations in either species, generally speaking, they differ only very little between highland and lowland populations. There is also no sign of parallel rearrangements between *H. erato* and *H. melpomene* (Fig. 4*A*). However, a specific rearrangement on *H. erato* chromosome 2 (Chr2) stands out. Here, among all *H. erato* samples, we observed unusually high beadTag sharing between windows 1.1 Mbp apart, in a manner indicative of an inversion (Fig. 4*B*; inferred junctions at left [Herato0204:172503] and right [Herato0204:1290057]). This elevated signal is especially strong among lowland *H. erato lativitta* butterflies and is distinct from previously reported inversions on chromosome 2 (25, 44, 45), suggesting that the inversion may have originated in *H. erato lativitta* or its close relatives. The two junctions bracket a region of elevated F_{ST} (Fig. 4*B*), suggesting reduced gene flow in this genomic region between the highland *H. erato notabilis* and lowland *H. erato lativitta* races. Using LR data, we have the opportunity to directly detect molecules that span the inverted Left and Right junctions (i.e., $Left_{out}$–$Right_{in}$ and $Left_{in}$–$Right_{out}$) (Fig. 4 *B*, *Left*, "bow-tie" pattern). This we find in 113 individuals homozygous for the inversion (as well as 163 heterozygotes and 152 individuals homozygous for the uninverted orientation; Fig. 4 *B*, *Middle*) and show that the inversion is segregating across the *H. erato* hybrid zone.

Haplotype Frequencies across the Hybrid Zone. In both species, migration and gene flow between the high- and lowland forms generated clines across the genome (i.e., gradient in gene frequencies

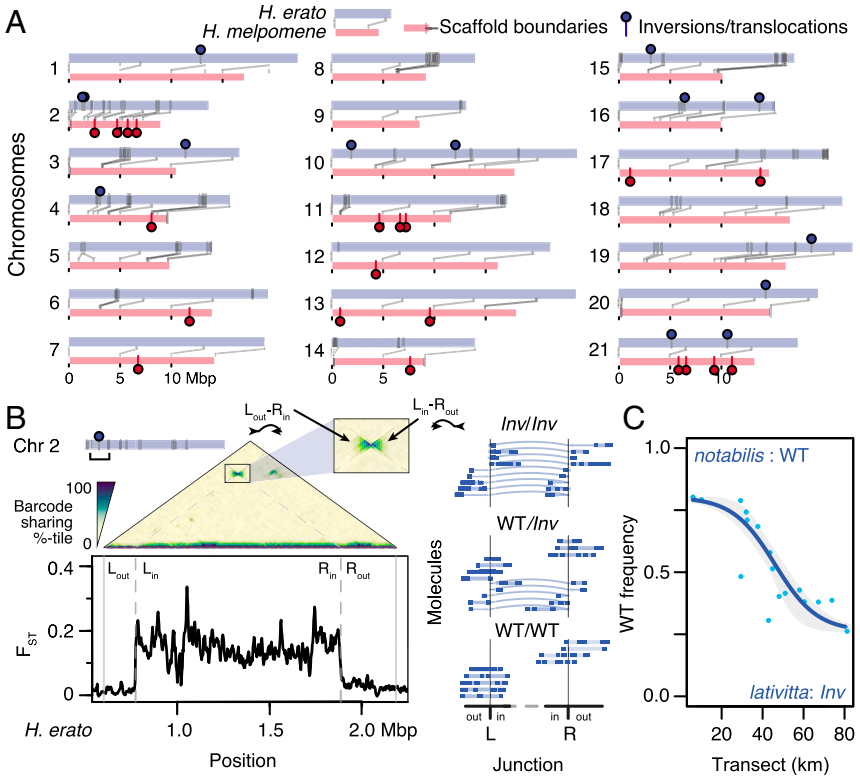

Fig. 4. Distinct structural rearrangements across the parallel hybrid zones. (*A*) Locations of major structural rearrangements (translocations and inversions) in the two *Heliconius* hybrid zones. Chromosome homologs are shown in pairs, with lines connecting syntenic positions between *H. erato* (gray) and *H. melpomene* (red; lines: dark gray bars mark scaffold boundaries; circles mark major inversions or translocations). In contrast to the parallelism at divergent peaks shown in Fig. 3, major structural rearrangements tend to be unique for each species. (*B*) Detection of a major inversion on *H. erato* Chr2. The average LR molecule spans multiple 10-kbp windows. Thus, the extent of beadTag sharing across windows (10 kbp here) can reveal discrepancies between the physical molecules and the reference assembly as well as across populations. The triangular matrix shows a heatmap of barcode sharing (color indicates genome-wide percentile) juxtaposed against genetic distance (F_{ST}) across the pure *notabilis* and *lativitta* races. Inversions appear as a "bow-tie"–shaped pattern across the inverted junction boundaries (L, left boundary of the inversion; R, right boundary of the inversion; out/in, outside or inside of the inversion; $Left_{in}$/$Right_{out}$ and $Left_{out}$/$Right_{in}$, zoomed inset). This inversion coincides with a plateau of elevated genetic distance across the *notabilis* and *lativitta* races. Dotted lines mark the inferred inversion boundaries at Herato0204:172503-1290057. Molecules from three individuals representing the three inversion collinear versus heterokaryotypes are shown (inferred inversion indicated with curved arrows). (*C*) The Chr2 inversion shows a clinal distribution across the *notabilis*–*lativitta* hybrid zone (frequency of wild-type [WT] karyotype: WT, blue dots; fitted cline: blue line; confidence interval: gray envelope).

along the transect) (*SI Appendix*, Fig. S13). For example, at the Chr2 inversion in *H. erato*, the wild-type orientation decreases from 80.2% in the highland *notabilis* race to 26.3% in the lowland *lativitta* one (or 73.7% inverted; estimated center of zone: 46.6 ± 3.2 km; width: 53.44 ± 23.7 km; Fig. 4*C*). In cline analysis, the steepness indicates selection: neutral loci introgress freely and produce wide and shallow clines, whereas strongly selected loci remain distinct between the races and produce sharp and narrow clines. Accordingly, the major color loci are among the narrowest clines in the genome (*SI Appendix*, Fig. S14). Plotting of both phenotypic and haplotype clines at *optix* (Chr18) and *WntA* (Chr10) in the two comimetic species shows a striking pattern: in each species, the *WntA* cline center is shifted east toward the lowlands (at a large drop in elevation between kilometers 45 and 50) relative to the *optix* cline or indeed much of the genome (Fig. 5*A*; 15.28 and 20.87 km in *H. erato* and *H. melpomene*, respectively; *SI Appendix*, Fig. S14). However, at these two color loci, both the positions and widths of the clines closely mirror each other between *H. erato* and *H. melpomene* (*optix*, centers: 31.9 versus 28.9 km; widths: 15.7 versus 15.2 km; *WntA*, centers: 47.1 versus 49.8 km; widths: 19.0 versus 24.8 km; Fig. 5 *A* and *B* and *SI Appendix*, Table S11). Interestingly, the minor color loci (*vvl* and *cortex*) track with a different major color locus in each species: *vvl* (Chr13) tracks with *WntA* in *H. erato* and *cortex* (Chr15) tracks with *optix* in *H. melpomene* (*SI Appendix*, Figs. S11 and S12). Both show broader cline widths likely due to dominance and reduced phenotypic effects, and hence weaker selection, because recessive alleles can introgress easily when dominant homozygotes are indistinguishable from heterozygotes (*SI Appendix*, Table S11; ref. 28). In fact, genetic crosses and association mapping suggest that the displacement of the clines at these minor loci may reflect the underlying genetics in refining the match across the comimics: *vvl* acts as a modifier for *WntA* in shaping the forewing band in *H. erato* butterflies (Fig. 3*B*), and its cline is shifted to more closely match this locus only in *H. erato* but not in *H. melpomene.* Likewise, *cortex* (or, possibly, the closely linked *domeless*/*washout*) modifies *optix* to generate the fully Dennis–Ray phenotype only in *H. melpomene*, and its cline shows a closer match with *optix* there than in *H. erato* (Fig. 3*C* and *SI Appendix*, Fig. S5 *F* and *H* and S15) (28).

The finding of displaced clines at the major color loci makes this Ecuadorean hybrid zone unique among *Heliconius* hybrid zones (n.b., this phenomenon is more common in other species, see refs. 46 and 47). In all other *Heliconius* hybrid zones, both *H. erato* and *H. melpomene* clines across different color loci coincide [e.g., Peru (48, 49) and Panama and Colombia (50)]. This may be because whenever clines overlap, they tend to be pulled together into coincidence due to increased LD (51). To better understand our observation of displaced clines, we estimated dispersal and dominance, from which we can derive an estimate of the strength of selection, all key parameters that govern cline dynamics.

Whenever clines overlap, LD is generated through the admixture of distinct populations, even between unlinked loci. Here, we do not expect LD between shifted clines, but neither do we see any significant association between coincident clines (*SI Appendix*, Tables

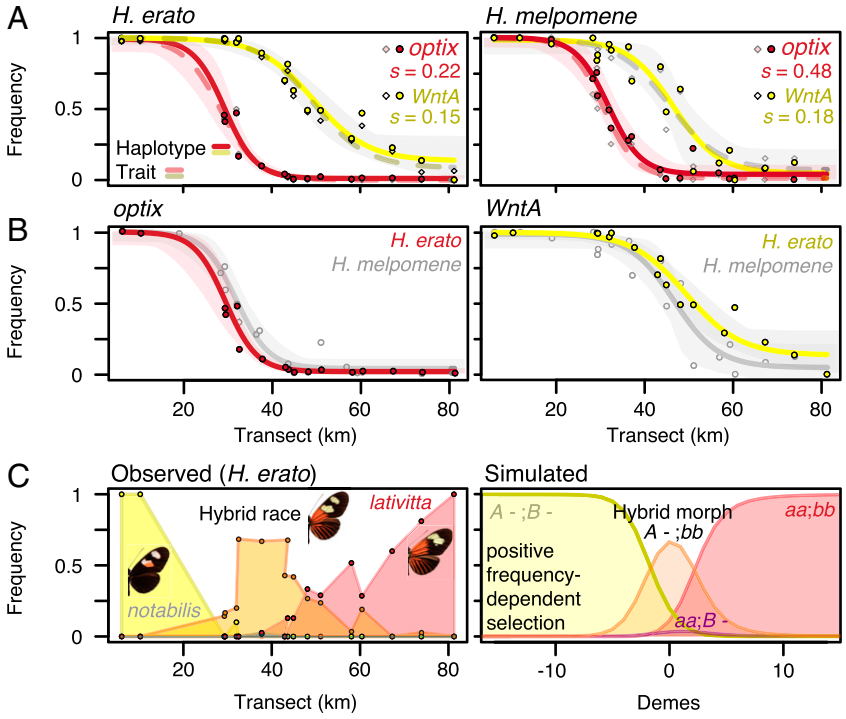

Fig. 5. Müllerian comimicry and the emergence of a hybrid race due to mirrored cline displacement of color traits. (*A*) Major color traits segregating across the Ecuadorian hybrid zones show a clinal distribution of haplotype frequencies along the transect in both *H. erato* (*Left*) and *H. melpomene* (*Right*). There is strong agreement in cline fits between haplotype frequencies (filled circles; cline: colored lines with 95% confidence envelope) and phenotype frequencies (diamonds and dashed lines). The gene *optix* (red) controls the red color pattern (see Fig. 3 and *SI Appendix*, Fig. S5 *E* and *F*) and shows a steeper and west-shifted cline compared to *WntA* (yellow), which controls the number of forewing bands (Fig. 3). (*B*) Clines are mirrored at both *optix* (*Left*) and *WntA* (*Right*) loci between *H. erato* (filled circles and colored lines) and its *H. melpomene* comimic (empty circles and gray lines). (*C*) Emergence of a novel hybrid morph in the middle of the hybrid zone. Due to the displaced clines, hybrid *H. erato* butterflies (*Left*; orange symbols and lines; middle wings) can display the highland *notabilis* double band (*Left*; yellow) along with the lowland *lativitta dennis* and *rays* (*Right*; red). This hybrid morph carries homozygous $WntA^{H/H}$ and $optix^{L/L}$ genotypes and is therefore true breeding. Simulation results show the frequencies of the four morphs, assuming complete dominance at two loci. Morph *i* has fitness $1 + s_i (P_i - Q_i)$, which increases linearly with its own frequency, P_i. Even when clines at the two loci start fully coincident, they can shift apart and produce displaced clines over time (here, generation 1,000), if there is a fitness advantage to one of the hybrid genotypes, here $s_{A\text{-};\,bb} = 0.25$, and the rest having $s_i = 0.1$.

S14 and S15; see *SI Appendix*, *SI Text* for details). Using these values, we can set upper limits to the dispersal rates (σ) that would be consistent with observed cline widths (w) and weak LD. These gave <3.2 km in *H. erato* and 6.2 km in *H. melpomene* (*SI Appendix*, Table S15; assuming linear frequency–dependent selection with complete dominance, so that $\sigma \sim w\sqrt{s/12}$, ref. 48). These upper limits to dispersal are slightly larger than, and therefore consistent with, previous estimates in Peru (48).

We thus also estimated the strength of selection (s) for each locus separately using the cline widths (w) using the relationship $s = 8(\sigma/w)^2$ (52) and previous—and consistent—estimates of dispersal rates (σ) of 2.6 km in *H. erato* and 3.7 km in *H. melpomene* (ref. 48; *SI Appendix*, Table S15). Unlike s estimates from LD across loci, which may be lower due to weakly selected loci, selection coefficients estimated from cline widths at each locus were very strong at *optix* (0.22 and 0.48 in *H. erato* and *H. melpomene*, respectively) and still considerable at *WntA* (0.15 and 0.18, respectively).

Emergence of a Novel Hybrid Morph. One consequence of the displaced *WntA* and *optix* clines is that a novel hybrid morph combining the highland double forewing band with the lowland Dennis–Ray pattern ($WntA^{H/H}$;$optix^{L/L}$) has become common in the middle of both *Heliconius* hybrid zones. Indeed, from 33 to 45 km along the transect, this $WntA^{H/H}$;$optix^{L/L}$ genotype is the most common morph in both species (Figs. 2*B* and 5 *C*, *Left* and *SI Appendix*, Tables S6 and S7). We used deterministic two-locus simulations to test whether positive frequency–dependent selection could maintain a hybrid morph. Our data with the shifted cline centers are largely consistent with dominance of the lowland over the highland allele at *WntA* and *optix* and fit with the higher similarity of F1 individuals to lowland rather than highland individuals. This implies introgression of highland alleles into the lowland populations with four possible morphs ($WntA^{H/H}$;$optix^{H/H}$, $WntA^{H/H}$;$optix^{L/-}$, $WntA^{L/-}$;$optix^{H/H}$, $WntA^{L/-}$;$optix^{L/-}$) because the highland alleles can introgress as "hidden" recessive alleles in the latter three morphs (*SI Appendix*, Table S16). Over time, the cline centers are expected to shift westward toward the highlands as a result of dominance at each locus (53). All else being equal for the four morphs, clines that are well displaced to start with can remain separate because each cline moves at the same speed. But if they overlap, either initially or because the leading cline stops moving due to other factors, LD pulls them together (*SI Appendix*, Fig. S16).

The above scenario will produce a hybrid morph that can persist, perhaps indefinitely, but its distribution will shift westward over time given suitable habitats. Since the distribution of the hybrid morph appears to be stable (54), there must be additional factors that cause clines (including ones that at first coincide) to shift and remain apart. These could include genetic incompatibilities or a selective advantage of one hybrid morph over the other (46, 47), but we favor a model in which the hybrid morph is favored or experiences stronger frequency-dependent selection, perhaps due to a more memorable phenotype for predators, which can maintain stable shifted clines as observed in the empirical data (*SI Appendix*, Fig. S16, *Bottom*). Unlike hybrid morphs with heterozygous genotypes, this $WntA^{H/H}$;$optix^{L/L}$ novel hybrid morph breeds true and has risen to appreciable frequencies, perhaps representing establishment of a hybrid race.

Discussion

The discovery and characterization of natural variation in the genome is a key first step in genetics and evolution. Such information can help us understand the genetic basis of trait variation and speciation. However, until now, it has not been easy to capture this variation as haplotypes in large population samples. Haplotagging solves this problem by generating linked-read data from hundreds of samples efficiently and affordably. While adding phasing information can be generally helpful for organisms over a wide range of LD, the power of haplotagging (and population haplotyping) is maximized in selective breeding or recently admixed populations. These data are far richer in information and permit the simultaneous characterization of both nucleotide and structural variation.

More broadly, this work highlights the advantage of combining broad population sampling with linkage information in large-scale LR data. Together, they allow efficient and accurate genome-wide haplotyping as opposed to genotyping. We hope that this work will spur development of improved algorithms (14) and experimental designs, such that future researchers may be able to perform (meta) genome assembly, phasing, imputation, and mapping in a single experiment. We anticipate that haplotagging or similar approaches (and eventually LR sequencing) may help drive the next phase of discoveries in model and nonmodel organisms alike.

We have used these data to demonstrate the early stages in establishment of a novel hybrid morph through the parallel displacement of clines in two species. In addition, we have also discovered >300 candidates under local or divergent selection, which opens up additional dimensions beyond wing color patterns to investigate this double hybrid zone. Somewhat to our surprise, our survey for structural rearrangements, such as inversions, found no consistent association with population differentiation in either species, suggesting that they do not play an early role in mediating divergence in the face of gene flow despite widespread support in the literature (40, 42, 55, 56) (reviewed in ref. 57, but see ref. 40). More broadly, our results further highlight the role of gene flow and local adaptation in the context of ecological speciation (38, 58, 59).

Materials and Methods

Animal Care and Use. All experimental procedures described in this study have been approved by the applicable ethics committee or authorities at the University of Calgary, Canada and Regierungspräsidium Tübingen, Germany. The study and collection of *Heliconius* butterflies has been approved by the Ecuadorian government. See *SI Appendix* for details.

Reference Genome Assemblies. Human: GRCh38; Mouse: mm10; *Heliconius* butterflies: *Heliconius erato demophoon* version 1 ("helera1_demo") and the *Heliconius melpomene melpomene* version 2.5 assembly (Hmel2.5).

Haplotagging Bead Assembly. Haplotagging beads are individually barcoded "Dynabeads M-280 Streptavidin magnetic beads" (Thermo Fisher Scientific) that are capable of DNA tagmentation through bead-immobilized active Tn5 transposase (*SI Appendix*, Fig. S1). The barcoding was achieved through extending the index sequences in Illumina's TruSeq adapter format to accommodate four combinatorial segments of a total of 26 nt. To assemble a haplotagging bead, adapters were bound onto microbeads and extended through ligation. Barcode diversity was achieved through a four-round split-and-pool aliquoting strategy in 96-well plates, one for each segment: "A," "B," "C," and "D." At the end of this procedure, each individual bead was coated with many adapter copies bearing an identical barcode for a total of 85 million unique combinations. These beads were then loaded with Tn5 transposase for library construction.

Sequencing Library Construction. The basic procedure of haplotagging library construction involved mixing genomic DNA with a pool of haplotagging beads and incubating for 10 min at 55 °C followed by Tn5 stripping (0.3% sodium dodecyl sulfate, 10 min at 55 °C) and PCR amplification (see *SI Appendix* for details). For single-plex libraries, 5 μl pooled haplotagging beads was used per 3 ng high molecular weight (HMW) DNA. For 96-plex multiplexing, 1.5 ng HMW DNA was tagged with 2.5 μl pooled beads and subsampled at a 1:10 (mice) or 1:24 ratio (butterflies) for pooling across samples. Amplified libraries were cleaned up and size-selected using Ampure magnetic beads (Beckman Coulter), Qubit quantified, and adjusted with 10 mM Tris, pH 8, 0.1 mM EDTA to 2.5 nM concentration for sequencing.

Sequencing and Demultiplexing. Pooled libraries were sequenced by a HiSeq 3000 (Illumina) instrument at the Genome Core Facility at the Max Planck Institute for Developmental Biology, typically with a 150+13+12+150 cycle run setting, such that the run produced 13 and 12 nt in the i7 and i5 index

reads, respectively. See *SI Appendix*, Fig. S2 and *SI Materials and* Methods for details on beadTag demultiplexing and handling. Code is available at https://github.com/evolgenomics/haplotagging.

Analysis and Phasing of Molecules. A schematic of the analytical pipeline is shown in *SI Appendix*, Fig. S2. Data analysis follows mainly standard short-read sequencing processing while retaining molecular information using a BX beadTag. For the human GM12878 and mouse F1(BL6xCAST) hybrid and backcross mice, known SNP sites were used to evaluate molecular accuracy and phasing performances using HAPCUT2 (10). For Longshanks mice, a set of variable sites from ref. 13 was used, and genotypes from the same individuals sequenced previously (13) were used to evaluate genotype imputation performance using STITCH (11).

Population Genomics of the Parallel *Heliconius* Hybrid Zones. Butterflies from the two species *Heliconius erato* ($n = 484$) and *Heliconius melpomene* ($n = 187$) from an overlapping hybrid zone in Ecuador were haplotagged and sequenced. See *SI Appendix*, Fig. S2C for details on the analytical pipeline. Briefly, the samples were analyzed for population structure (population clusters and F_{ST}) and signatures of selection, both for changes in allele frequencies (π) and haplotype lengths (iHS, ω). Additionally, a GWAS was performed to identify loci associated with wing color pattern traits. See *SI Appendix* for details.

Cline Analysis. One-dimensional analyses of cline position and shape were carried out using a subset of 35 transect sites with 928 *H. erato* and 440 *H. melpomene* butterflies (484 and 187 sequenced, respectively). For each trait (wing color pattern) or locus, changes in allele frequencies across the cline were analyzed to determine the best-fitting cline model, typically with two parameters, cline centers and widths. For the major color loci *WntA*, *optix*, *vvl*, and *cortex*, dominance was estimated and a deterministic two-locus model assuming positive linear frequency–dependent selection was used to simulate cline movements under a range of plausible starting conditions and selection parameters. The code is available at https://github.com/evolgenomics/haplotagging.

Data Availability. Analysis codes are available at the following repositories: https://github.com/evolgenomics/haplotagging and https://github.com/evolgenomics/HeliconiusHaplotagging. Sequence data have been deposited at the National Center for Biotechnology Information Sequence Read Archive under the BioProject accession number PRJNA670070. Phenotypic and collection data for all *Heliconius* butterfly specimens are given in Dataset S6.

ACKNOWLEDGMENTS. We thank Felicity Jones for input into experimental design, helpful discussion, and improving the manuscript. We thank the C.R., C.D.J., Y.F.C., and Jones laboratory members for support, insightful scientific discussion, and improving the manuscript. We thank the C.R. laboratory members, the Animal Resource Centre staff at the University of Calgary, and Caroline Schmid and Ann-Katrin Geysel at the Friedrich Miescher Laboratory for animal husbandry. We thank Christa Lanz, Rebecca Schwab, and Ilja Bezrukov for assistance with high-throughput sequencing and associated data processing and Andre Noll and the MPI Tübingen information technology team for computational support. We thank Ben Haller and Richard Durbin for helpful discussions. We thank David M. Kingsley, James Mallet, the Editor, and three anonymous reviewers for thoughtful input that has greatly improved our manuscript. J.I.M. is supported by a Research Fellowship from St. John's College, Cambridge, and a Branco Weiss Fellowship. A.D. was supported by European Research Council Consolidator Grant 617279 "EvolRecombAdapt," under Dr. Felicity Jones. C.R. is supported by the Natural Sciences and Engineering Research Council of Canada Discovery Grant No. 4181932 and by the Faculty of Veterinary Medicine at the University of Calgary. C.D.J. is supported by BBSRC Grant BB/R007500 and European Research Council Advanced Grant 339873 "SpeciationGenetics." M.K. and Y.F.C. are supported by the Max Planck Society and European Research Council Starting Grant 639096 "HybridMiX."

1. N. H. Barton, P. D. Keightley, Understanding quantitative genetic variation. *Nat. Rev. Genet.* **3**, 11–21 (2002).
2. O. Seehausen *et al.*, Genomics and the origin of species. *Nat. Rev. Genet.* **15**, 176–192(2014).
3. G. Sella, N. H. Barton, Thinking about the evolution of complex traits in the era of genome-wide association studies. *Annu. Rev. Genomics Hum. Genet.* **20**, 461–493(2019).
4. R. Tewhey, V. Bansal, A. Torkamani, E. J. Topol, N. J. Schork, The importance of phase information for human genomics. *Nat. Rev. Genet.* **12**, 215–223(2011).
5. N. R. Garud, P. W. Messer, E. O. Buzbas, D. A. Petrov, Recent selective sweeps in North American *Drosophila melanogaster* show signatures of soft sweeps. *PLoS Genet.* **11**, e1005004(2015).
6. S. Amini *et al.*, Haplotype-resolved whole-genome sequencing by contiguity-preserving transposition and combinatorial indexing. *Nat. Genet.* **46**, 1343–1349(2014).
7. G. X. Y. Zheng *et al.*, Haplotyping germline and cancer genomes with high-throughput linked-read sequencing. *Nat. Biotechnol.* **34**, 303–311(2016).
8. F. Zhang *et al.*, Haplotype phasing of whole human genomes using bead-based barcode partitioning in a single tube. *Nat. Biotechnol.* **35**, 852–857(2017).
9. O. Wang *et al.*, Efficient and unique cobarcoding of second-generation sequencing reads from long DNA molecules enabling cost-effective and accurate sequencing, haplotyping, and de novo assembly. *Genome Res.* **29**, 798–808(2019).
10. P. Edge, V. Bafna, V. Bansal, HapCUT2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Res.* **27**, 801–812(2017).
11. R. W. Davies, J. Flint, S. Myers, R. Mott, Rapid genotype imputation from sequence without reference panels. *Nat. Genet.* **48**, 965–969(2016).
12. M. Marchini *et al.*, Impacts of genetic correlation on the independent evolution of body mass and skeletal size in mammals. *BMC Evol. Biol.* **14**, 258(2014).
13. J. P. Castro *et al.*, An integrative genomic analysis of the Longshanks selection experiment for longer limbs in mice. *eLife* **8**, e42014 (2019).
14. R. W. Davies *et al.*, Rapid genotype imputation from sequence with reference panels. *Nat. Genet.* **48**, 965–969 (2016).
15. The 1000 Genomes Project Consortium. A map of human genome variation from population-scale sequencing. *Nature* **467**, 1061–1073 (2010).
16. B. Pasaniuc *et al.*, Extremely low-coverage sequencing and imputation increases power for genome-wide association studies. *Nat. Genet.* **44**, 631–635 (2012).
17. C. D. Jiggins, "What can we learn about adaptation from the wing pattern genetics of *Heliconius* butterflies?" in *Diversity and Evolution of Butterfly Wing Patterns: An Integrative Approach*, T. Sekimura, H. F. Nijhout, Eds. (Springer Singapore, Singapore, 2017), pp. 173–188.
18. J. Mallet, M. Joron, Evolution of diversity in warning color and mimicry: Polymorphisms, shifting balance, and speciation. *Annu. Rev. Ecol. Syst.* **30**, 201–233 (1999).
19. G. M. Langham, Specialized avian predators repeatedly attack novel color morphs of Heliconius butterflies. *Evolution* **58**, 2783–2787 (2004).
20. J. Mallet, N. H. Barton, Strong natural selection in a warning-color hybrid zone. *Evolution* **43**, 421–431 (1989).
21. M. Joron *et al.*, Chromosomal rearrangements maintain a polymorphic supergene controlling butterfly mimicry. *Nature* **477**, 203–206 (2011).
22. R. D. Reed *et al.*, *optix* drives the repeated convergent evolution of butterfly wing pattern mimicry. *Science* **333**, 1137–1141 (2011).
23. A. Martin *et al.*, Diversification of complex butterfly wing patterns by repeated regulatory evolution of a *Wnt* ligand. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 12632–12637 (2012).
24. N. J. Nadeau *et al.*, The gene cortex controls mimicry and crypsis in butterflies and moths. *Nature* **534**, 106–110 (2016).
25. S. M. Van Belleghem *et al.*, Complex modular architecture around a simple toolkit of wing pattern genes. *Nat. Ecol. Evol.* **1**, 52 (2017).
26. J. J. Lewis *et al.*, Parallel evolution of ancient, pleiotropic enhancers underlies butterfly wing pattern mimicry. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 24174–24183 (2019).
27. K. M. Kozak *et al.*, Multilocus species trees show the recent adaptive radiation of the mimetic *heliconius* butterflies. *Syst. Biol.* **64**, 505–524 (2015).
28. P. A. Salazar Carrión, "Hybridization and the genetics of wing colour-pattern diversity in Heliconius butterflies," PhD thesis, University of Cambridge, Cambridge, UK (2013).
29. N. J. Nadeau *et al.*, Population genomics of parallel hybrid zones in the mimetic butterflies, *H. melpomene* and *H. erato*. *Genome Res.* **24**, 1316–1333 (2014).
30. C. A. Buerkle, C. Lexer, Admixture as the basis for genetic mapping. *Trends Ecol. Evol.* **23**, 686–694 (2008).
31. N. H. Barton, G. M. Hewitt, "Evolution and speciation" in *Hybrid Zones and Speciation*, W. R. Atchley, D. S. Woodruff, Eds. (Cambridge University Press, NY, 1981), pp. 109–145.
32. M. Joron *et al.*, A conserved supergene locus controls colour pattern diversity in Heliconius butterflies. *PLoS Biol.* **4**, e303 (2006).
33. L. Ferguson *et al.*, Characterization of a hotspot for mimicry: Assembly of a butterfly wing transcriptome to genomic sequence at the HmYb/Sb locus. *Mol. Ecol.* **19** (suppl. 1), 240–254 (2010).
34. S. W. Baxter *et al.*, Convergent evolution in the genetic basis of Müllerian mimicry in heliconius butterflies. *Genetics* **180**, 1567–1577 (2008).
35. L. Livraghi *et al.*, The gene cortex controls scale colour identity in Heliconius. *bioRxiv* [Preprint] (2020). 10.1101/2020.05.26.116533.
36. Y. Kim, R. Nielsen, Linkage disequilibrium as a signature of selective sweeps. *Genetics* **167**, 1513–1524 (2004).
37. B. F. Voight, S. Kudaravalli, X. Wen, J. K. Pritchard, A map of recent positive selection in the human genome. *PLoS Biol.* **4**, e72 (2006).
38. R. W. R. Wallbank *et al.*, Evolutionary novelty in a butterfly wing pattern through enhancer shuffling. *PLoS Biol.* **14**, e1002353 (2016).
39. J. Hanly, "Developmental basis of wing pattern diversity in Heliconius butterflies," PhD thesis, University of Cambridge, Cambridge, UK (2017).
40. J. L. Feder, J. B. Roethele, K. Filchak, J. Niedbalski, J. Romero-Severson, Evidence for inversion polymorphism related to sympatric host race formation in the apple maggot fly, *Rhagoletis pomonella*. *Genetics* **163**, 939–953 (2003).
41. D. B. Lowry, J. H. Willis, A widespread chromosomal inversion polymorphism contributes to a major life-history transition, local adaptation, and reproductive isolation. *PLoS Biol.* **8**, e1000500 (2010).

42. F. C. Jones *et al*.; Broad Institute Genome Sequencing Platform & Whole Genome Assembly Team, The genomic basis of adaptive evolution in threespine sticklebacks. *Nature* **484**, 55–61 (2012).
43. A. Dréau, V. Venu, E. Avdievich, L. Gaspar, F. C. Jones, Genome-wide recombination map construction from single individuals using linked-read sequencing. *Nat. Commun.* **10**, 4309 (2019).
44. J. W. Davey *et al*., No evidence for maintenance of a sympatric *Heliconius* species barrier by chromosomal inversions. *Evol. Lett.* **1**, 138–154 (2017).
45. N. B. Edelman *et al*., Genomic architecture and introgression shape a butterfly radiation. *Science* **366**, 594–599 (2019).
46. T. Hatfield, N. Barton, J. B. Searle, A model of a hybrid zone between two chromosomal races of the common shrew (*Sorex araneus*). *Evolution* **46**, 1129–1145 (1992).
47. S. R. Virdee, G. M. Hewitt, Clines for hybrid dysfunction in a grasshopper hybrid zone. *Evolution* **48**, 392–407 (1994).
48. J. Mallet *et al*., Estimates of selection and gene flow from measures of cline width and linkage disequilibrium in *heliconius* hybrid zones. *Genetics* **124**, 921–936 (1990).
49. N. Rosser, K. K. Dasmahapatra, J. Mallet, Stable *Heliconius* butterfly hybrid zones are correlated with a local rainfall peak at the edge of the Amazon basin. *Evolution* **68**, 3470–3484 (2014).
50. E. V. Curran *et al*., Müllerian mimicry of a quantitative trait despite contrasting levels of genomic divergence and selection. *Mol. Ecol.* **29**, 2016–2030 (2020).
51. J. M. Szymura, N. H. Barton, Genetic analysis of a hybrid zone between the fire-bellied toads, *Bombina bombina* and *B. variegta*, near Cracow in Southern Poland. *Evolution* **40**, 1141–1159 (1986).
52. J. Mallet, N. Barton, Inference from clines stabilized by frequency-dependent selection. *Genetics* **122**, 967–976 (1989).
53. J. Mallet, Hybrid zones of *Heliconius* butterflies in Panama and the stability and movement of warning colour clines. *Heredity* **56**, 191–202 (1986).
54. M. G. Emsley, The geographical distribution of the color-pattern components of *Heliconius erato* and *Heliconius melpomene* with genetical evidence for the systematic relationship between the two species. *Zoologica (N.Y.)* **49**, 245–286 (1965).
55. J. Kitano *et al*., A role for a neo-sex chromosome in stickleback speciation. *Nature* **461**, 1079–1083 (2009).
56. R. Faria *et al*., Multiple chromosomal rearrangements in a hybrid zone between *Littorina saxatilis* ecotypes. *Mol. Ecol.* **28**, 1375–1393 (2019).
57. M. Wellenreuther, L. Bernatchez, Eco-evolutionary genomics of chromosomal inversions. *Trends Ecol. Evol.* **33**, 427–440 (2018).
58. D. A. Marques, J. I. Meier, O. Seehausen, A combinatorial view on speciation and adaptive radiation. *Trends Ecol. Evol.* **34**, 531–544 (2019).
59. J. Mavárez *et al*., Speciation by hybridization in *Heliconius* butterflies. *Nature* **441**, 868–871 (2006).